

ODLS 2016

Workshop on Ontologies and Data in Life Sciences,
Sep 29-30, Halle (Saale), Germany

Ontological interpretation of biomedical database annotations

Filipe Santana da Silva¹, Ludger Jansen²,
Fred Freitas¹, Stefan Schulz³

¹Centro de Informática (CIn), Universidade Federal de Pernambuco (UFPE), Recife, Brazil

²Institut für Philosophie, Universität Rostock, Germany

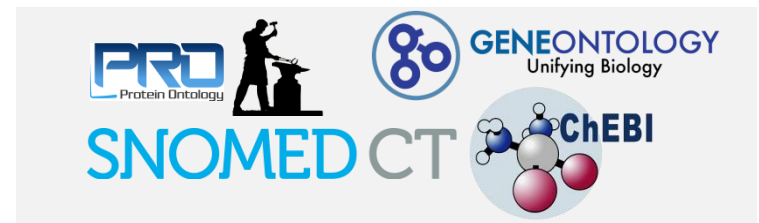
³Institut für Medizinische Informatik, Statistik und Dokumentation, Medizinische Universität
Graz, Austria

Biological Databases and Bio-Ontologies

- Two worlds: Bio-DBs



Bio-ontologies



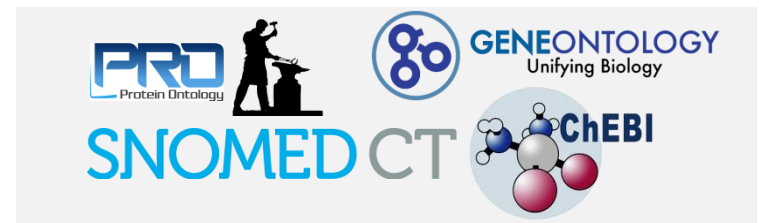
- How are they related to each other?
- Can their content be expressed by a unified model of meaning?
- Is database content of ontological nature?
- Can OWL be used as a language to express
 - bio-ontology content
 - bio-database structure
 - bio-database content
- Which is the added value?

Biological Databases and Bio-Ontologies

■ Two worlds: Bio-DBs



Bio-ontologies



- Store summarized results of laboratory experiments
- Classical database structure
- Values:
 - Numeric
 - Textual
 - Symbolic (codes from ontologies)

- Provide definitions
- Provide axioms that are universally true
- Obey formal semantics
- Main use case:
 - annotation of biological database entries



Example – tabular Bio-DB structure



ID	PR Protein	PR Gene	Organism (NCBI Tax)	GO Biological Process	GO Molecular Function	GO Cellular component	Ensembl ID	Ensembl Phenotype
F1MEW4	CBS	CBS	<i>Bos taurus</i>	blood vessel remodelling;...	Cystathionine beta-synthase;...	cytoplasm;...	ENSBTAT00000000184; ...	No phenotype associated
Q99707	MS	MS	<i>Homo sapiens</i>	cobalamin metabolic process;...	cobalamin binding;...	cytoplasm;...	ENST000000366577; ENST000000533889	Neural tube defect.; Megaloblastic anemia;...
F1RF82	MTHFR	MTHFR	<i>Sus scrofa</i>	homocysteine metabolic process;...	modified amino acid binding;...	cytosol	ENSSSCT00000003805	No phenotype associated
Q93088	BHMT	BHMT	<i>Homo sapiens</i>	amino acid betaine catabolic process;...	zinc ion binding;...	protein complex;...	ENST00000274353	Liver tumor; Coronary artery disease;...

Example – tabular Bio-DB structure



ID	PR Protein	PR Gene	Organism (NCBI Tax)	GO Biological Process	GO Molecular Function	GO Cellular component	Ensembl ID	Ensembl Phenotype
F1MEW4	CBS	CBS	<i>Bos taurus</i>	blood vessel remodelling;...	Cystathionine beta-synthase;...	cytoplasm;...	ENSBTAT00000000184; ...	No phenotype associated
Q99707	MS	MS	<i>Homo sapiens</i>	cobalamin metabolic process;...	cobalamin binding;...	cytoplasm;...	ENST000000366577; ENST000000533889	Neural tube defect.; Megaloblastic anemia;...
F1RF82	MTHFR	MTHFR	<i>Sus scrofa</i>	homocysteine metabolic process;...	modified amino acid binding;...	cytosol	ENSSSCT00000003805	No phenotype associated
Q93088	BHMT	BHMT	<i>Homo sapiens</i>	amino acid betaine catabolic process;...	zinc ion binding;...	protein complex;...	ENST00000274353	Liver tumor; Coronary artery disease;...

'is included in'

'is included in'

'is included in'

'is participant in'

'is part of'

'includes'

Example – tabular Bio-DB structure

more abstract:

- a database record informs about experimental evidence that:
 - Proteins of the type *Prot1*
 - participate in Processes of type $BProc_1 \dots BProc_k$ within organisms of type Org_1
 - are active in cellular components of type $CComp_1$ or $CComp_2$ or ... $CComp_x$ within organisms of type Org_1
 - participate in Processes that have small molecules of type $Mol_1, Mol_2 \dots Mol_y$ as outcome
 - within organisms of type Org_1 – if dysfunctional – Org_1 has dispositions to develop the phenotypes (disorders) $Phen_1 \dots Phen_z$

This information is not explicitly contained in the database – it is implicitly shared by database users and curators

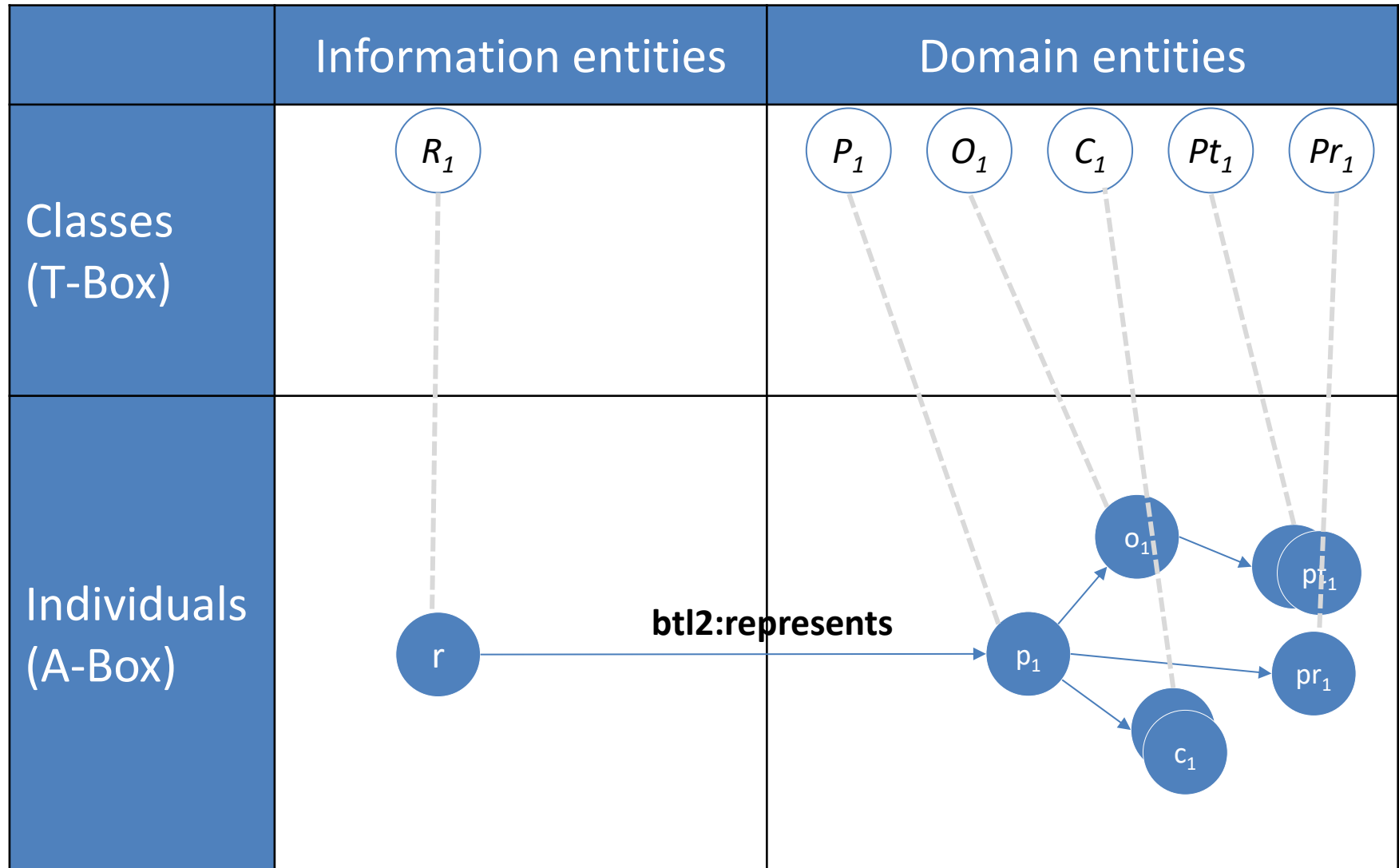
Ontological framework

	Information entities	Domain entities
Classes (T-Box)	• Bio-DB • Database record • Data item	• <i>Homo sapiens</i> • <i>Megaloblastic anemia</i> • <i>Cobalamin binding</i> • <i>Methionin synthase</i> • (...)
Individuals (A-Box)	• Uniprot • Ensembl • Database record about Methionin Synthase in Homo Sapiens • Data item, such as "cobalamin binding" in this record	• John Doe, of which tissue is stored in a biobank and analysed in a lab • John's megaloblastic anemia • A cobalamin binding process observed in the lab within a tissue sample from John • a dysfunctional Methionin synthase protein molecule in John's tissue

Denotation

	Information entities	Domain entities
Classes (T-Box)	- Bio-DB - Database record - Data item	<i>Homo sapiens</i> <i>Megaloblastic anemia</i> <i>Cobalamin binding</i> <i>Methionin synthase</i> (...)
Individuals (A-Box)	- UniProt - Ensembl - Database record about Methionin Synthase in Homo Sapiens - Data item, such as "cobalamin binding" in	- John Doe, of which tissue is stored in a biobank and analysed in a lab - John's megaloblastic anemia - A cobalamin binding process observed in the lab within a tissue sample from John - a dysfunctional Methionin synthase protein molecule in John's tissue

Case 1: database entry represents individuals



Multiple defined subclasses

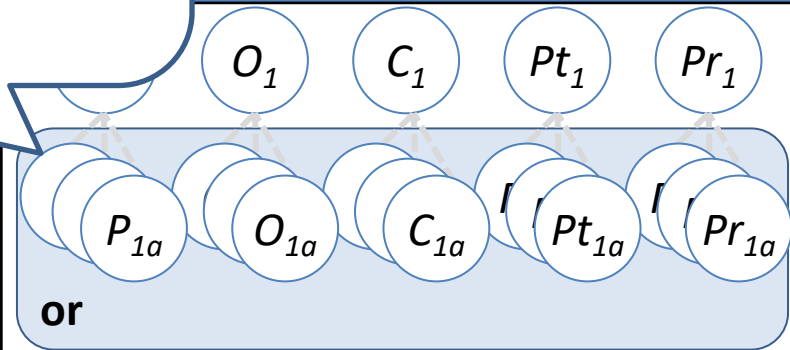
$\text{'Prot}_i \text{ Dysf in Org}_{i1} \text{ with Phen}_{i1, \dots, il}$ equivalentTo
 $\text{'Prot}_i \text{ Dysf in Org}_{i1}$ and **'is part of'** some $(\text{Org}_{i1} \text{ and includes some Phen}_{i1, \dots, im})$
 $\text{'Prot}_i \text{ in Org}_{i1} \text{ in BProc}_{i1, \dots, il}$ equivalentTo $\text{'Prot}_i \text{ in Org}_{i1}$ and **'is participant in'** some $\text{BProc}_{i1, \dots, im}$
 $\text{'Prot}_i \text{ in Org}_{i1} \text{ in CComp}_{i1, \dots, il}$ equivalentTo $\text{'Prot}_i \text{ in Org}_{i1}$ and **'is included in'** some $\text{CComp}_{i1, \dots, im}$
 $\text{'Prot}_i \text{ in Org}_{i1} \text{ with Mol}_{i1, \dots, il}$ equivalentTo
 $\text{'Prot}_i \text{ in Org}_{i1}$ and **'is participant in'** some $(\text{Process and 'has participant' some Mol}_{i1, \dots, im})$

$\text{'Org}_{i1} \text{ with Prot}_i$ equivalentTo Org_{i1} and **'has part'** some Prot_i
 $\text{'Org}_{i1} \text{ with Prot}_i \text{ Dysf}$ equivalentTo Org_{i1} and **'has part'** some $\text{'Prot}_i \text{ Dysf'}$
 $\text{'Org}_{i1} \text{ with Phen}_{i1, \dots, im} \text{ and Prot}_i \text{ Dysf}$ equivalentTo
 $\text{'Org}_{i1} \text{ with Prot}_i \text{ Dysf'}$ and **includes** some $\text{Phen}_{i1, \dots, il}$

represents classes

Classes
(T-Box)

Domain entities



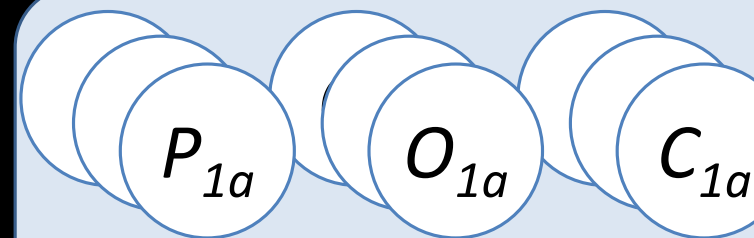
btl2:represents rdf:type only

Individuals
(A-Box)

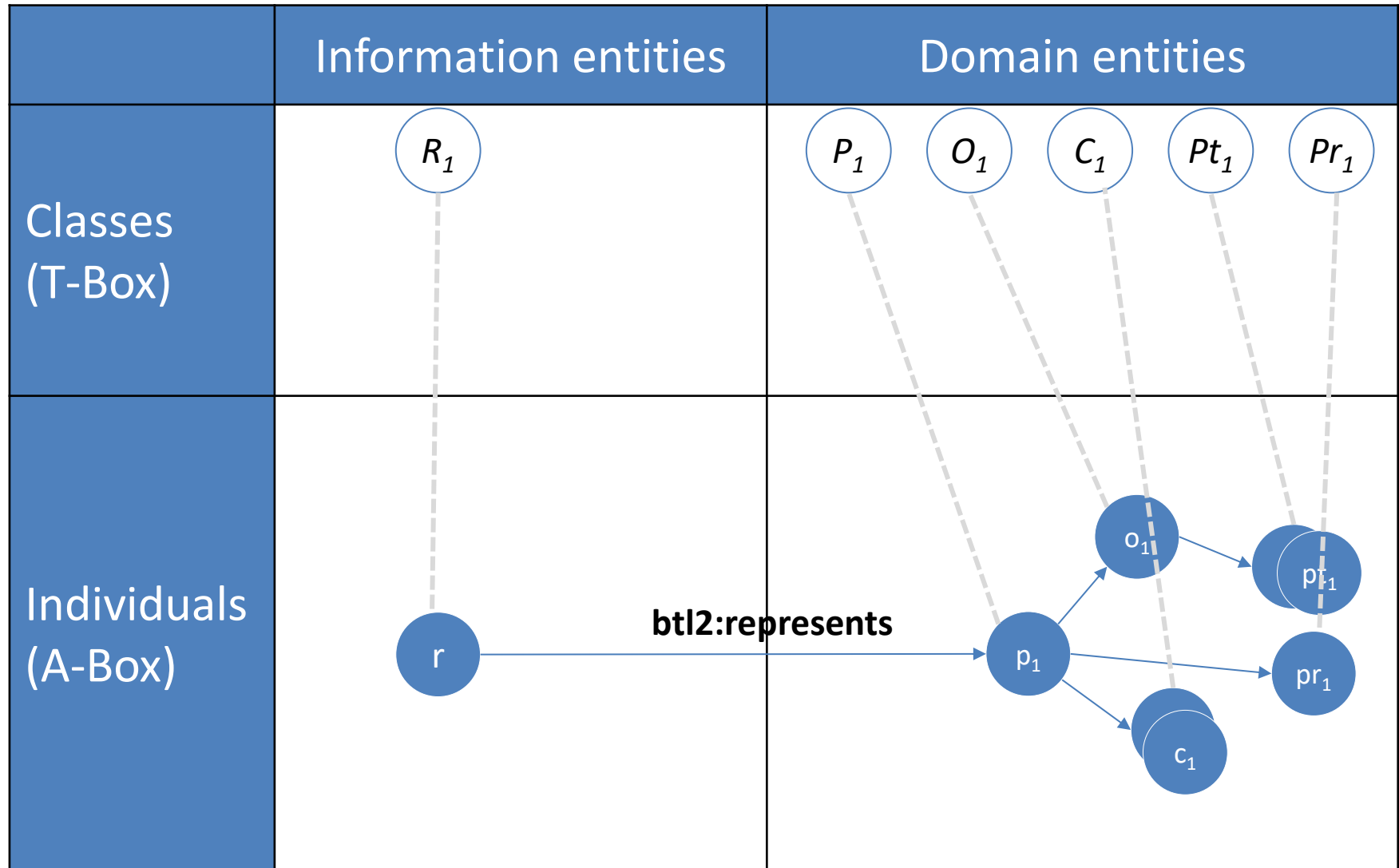


Multiple defined subclasses

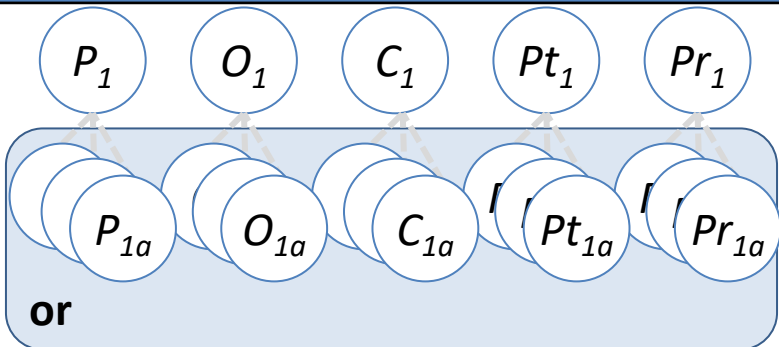
- ' $Prot_i$ Dysf in Org_{i1} with $Phen_{i1,...,il}$ ' equivalentTo
 - ' $Prot_i$ Dysf in Org_{i1} and **'is part of'** some (Org_{i1} and **includes** some $Phen_{i1,...,im}$)
- ' $Prot_i$ in Org_{i1} in $BProc_{i1,...,il}$ ' equivalentTo ' $Prot_i$ in Org_{i1} ' and **'is participant in'** some $BProc_{i1,...,im}$
- ' $Prot_i$ in Org_{i1} in $CComp_{i1,...,il}$ ' equivalentTo ' $Prot_i$ in Org_{i1} and **'is included in'** some $CComp_{i1,...,im}$
- ' $Prot_i$ in Org_{i1} with $Mol_{i1,...,il}$ ' equivalentTo
 - ' $Prot_i$ in Org_{i1} ' and **'is participant in'** some ($Process$ and **'has participant'** some $Mol_{i1,...,im}$)
- ' Org_{i1} with $Prot_i$ ' equivalentTo Org_{i1} and **'has part'** some $Prot_i$
- ' Org_{i1} with $Prot_i$ Dysf' equivalentTo Org_{i1} and **'has part'** some ' $Prot_i$ Dysf'
- ' Org_{i1} with $Phen_{i1,...,im}$ and $Prot_i$ Dysf' equivalentTo
 - ' Org_{i1} with $Prot_i$ Dysf' and **includes** some $Phen_{i1,...,il}$



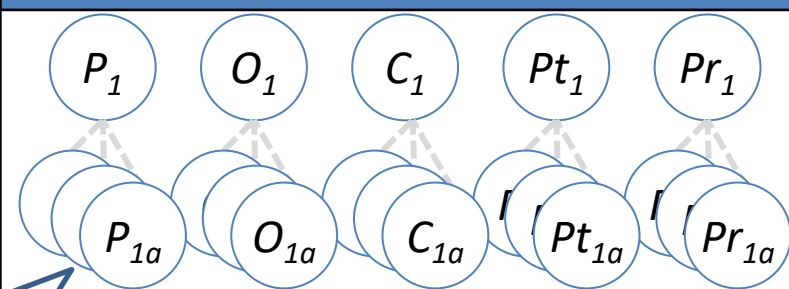
Querying: A-box query for individuals



Querying: T-box query for subclasses

	Information entities	Domain entities
Classes		 or
Individuals		

Querying: T-box query for subclasses

	Information entities	Domain entities
Classes		
Two step DL query: • Does subclass exist? • if not: no database entry else: determine superclass from bioontology		

Competency questions

(Q1) Which biological processes have proteins of the kind $Prot_i$ as participants?

$BProc_1$ and ('**has participant**' some $Prot_i$)

(Q2) In which cellular locations is $Prot_i$ active in organisms of the type Org_1 ?

'*Cellular component*' and ('**is included in**' some Org_1)
and (**includes** some $Prot_i$)

(Q3) Which proteins are involved in processes of the type $BProc$ in organisms of the type Org_1 ?

Protein and ('**is participant in**' some $BProc_1$)
and ('**is included in**' some Org_1)

Evaluation (one database record)

Model	Q1	Q2	Q3	Classes	Individuals	Axioms/ Assertions
A-Box	bp1001, bp2001, bp3001	cc1001, cc2001, cc3001	p1004	24	51	207
T-Box	$BProc_1$	$CComp_1$	$Prot_i$	68	0	149

Discussion (I)

- Both modelling solutions:
 - highly productive
 - scaling problems to be expected
- A-Box solution (prototypical individuals):
 - A-box reasoning more costly
 - Makes existential assumptions
- T-Box solution (multiple subclasses)
 - Theoretically allows non-referential entries
 - Simplified model: EL++

Discussion (II)

- Do biological database refer to ontological content?
 - No "real" universal statements on biological entities
 - Even no existential assumption
 - Dispositional statements to be discussed (see paper)
- Exercise best described as ontological representation of referring individuals
- Possible use case: non-disruptive querying of Bio-DBs where axioms of the annotation ontologies need to be explored

Conclusion

- Four ontological approaches - IND, SUBC, DISP and HYB
 - Structure and content of BIO-DBs
- Solution:
 - Expressiveness, DB retrieval and retrieval based on DL queries
 - Interpretation:
 - Denoted entities as prototypical individuals
 - Creation of defined subclasses
 - Database content as reporting dispositions

Funding: This work was funded by *Conselho Nacional de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES) 3914/2014-03 and *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq) 140698/2012-4.